

# Data Mining

## 1. Introduction

Nicolas Pasquier

<http://www.i3s.unice.fr/~pasquier>

1

## Terminologie

- Data Mining ou :
  - Fouille de données
  - Knowledge Discovery in Databases (KDD)
  - Extraction de connaissances à partir de données (ECD)
- Bases de Données Décisionnelles
  - Fouille de Données
    - Data Mining
  - Entrepôts de Données
    - Data Warehousing

2

## Motivation

- Explosion du volume des données
  - Outils et technologie de stockage performants
  - Recueil de données volumineux (transactions de ventes, cartes de crédit, images, etc.) : giga et tera-octets
- Exemple de Wal-Mart avec 20 millions de transactions, 483 processeurs parallèles
- Nécessité d'en tirer des connaissances utiles

3

## Définition

- Un processus non trivial d'extraction de modèles valides, nouveaux, potentiellement utiles et compréhensibles à partir de grands volumes de données
- Objectifs
  - Compréhension des données et des phénomènes sous-jacents (liens, récurrences, etc.)
  - Extrapolation d'informations pour la prédiction d'événements
  - Construction de modèles (calculs) pour la prédiction de valeurs (données)

4

## Domaines d'application

- Marketing
  - CRM (Customer Relationship Management), ventes croisées, Segmentation des marchés
    - Quels types de clients achètent quels types de produits?
    - Y-a-t-il des liens de causalité entre l'achat d'un produit P et d'un autre produit P'?
    - Quel est le comportement des clients au cours du temps?
  - Utiliser des données recueillies pour un produit similaire
    - Chercher des associations/corrélations entre produits
    - Chercher des segments dans les données décrivant les clients

5

## Domaines d'application

- Analyse et gestion des risques, Détection de fraudes
- Assurance, Domaine Bancaire (cartes de crédit, accord de crédit), Télécommunications
  - Peut-on caractériser les assurés qui font des déclarations d'accident frauduleuses?
  - Peut-on détecter un groupe de patients et un réseau de médecins qui ont des comportements anormaux
  - Quels sont les clients "à risque" pour l'accord de crédit?

6

## Domaines d'application

- Santé, Médecine
  - Aide au diagnostic
  - Étude de l'influence de certaines médications sur l'évolution d'une maladie
  - Recherche des médicaments les plus efficaces
- Biologie
  - Identifier des similarités dans des séquences d'ADN
  - Identifier les fonctions des gènes
- Astronomie
  - Identifier le type d'un objet à partir d'images stellaires données en pixels

7

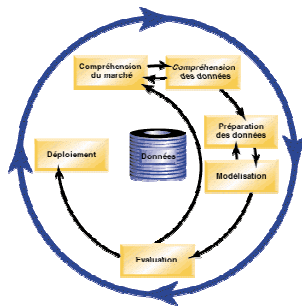
## Domaines d'application

- Télécommunications
  - Comment éviter l'attrition?
- Web Mining
  - Etudier le contenu, la structure ou l'usage des pages web
- Text mining (news group, email, documents)
  - Routage de courrier, rapports d'activité, etc.
- E-learning

8

## Méthodologie CRISP-DM

- Modèle de cycle de vie à six phases
- Arcs : dépendances les plus importantes et fréquentes
- Séquence non stricte : dans la plupart des projets, on passe d'une phase à l'autre en fonction des besoins



9

## Étapes du processus

- Comprendre le problème
  - Connaissance du domaine, buts poursuivis, données disponibles, déploiement des résultats
- Explorer : visualiser, questionner
- Créer la table de données (jeu de données)
  - Nettoyage et Intégration (60% du travail)
  - Réduction et Transformation
- Choisir la ou les fonctionnalités
  - Description, classification, régression, clustering, extraction d'associations, séries chronologiques

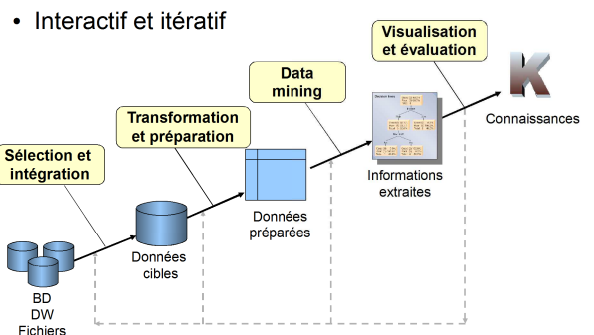
10

## Étapes du processus

- Choisir la (les) méthodes (algorithmes)
- Effectuer l'extraction
  - Recherche des modèles intéressants
- Évaluation du modèle
- Présentation des résultats

11

## Étapes du processus



12

## Data Mining vs. Statistiques

- Data Mining : intégration de techniques issues de divers domaines (bases de données, statistiques, apprentissage automatique, analyse de données, visualisation, etc.)
- Data Mining vs. Statistiques
  - Découvrir plutôt que vérifier
- Data Mining vs. Machine Learning
  - Manipuler des BD volumineuses plutôt que de « petits » ensembles d'apprentissage

13

## Principaux logiciels

- Logiciels « stand-alone »
  - SPSS Clementine
  - Salford Systems CART
  - Weka, RapidMiner, Orange, Tanagra (logiciels libres)
- Outils intégrés
  - IBM Intelligent Miner (IBM DB2)
  - SAS Enterprise Miner (SAS)
  - Oracle Data Mining (Oracle 10g)
  - DB Miner (IBM DB2)
  - SQL Server (Microsoft)

14

## Interfaces logicielles

- Logiciels utilisant la méthode CRISP-DM
  - Représentations graphique des traitements
  - Graphes des flux de données et traitements enchaînés
  - Définition de « run-times » pour automatiser les traitements
  - Présentation des résultats sous forme de rapports
- Logiciels ou modules intégrés aux SGBD
  - Requêtes de Data Mining
  - Stockage relationnel des données
  - Langages SQL étendu, DMQL, opérateurs spécifiques
  - Modules de génération de rapports du SGBD
  - Interface avec les modules de data warehousing

15

## Types de données

- Origine des données
  - BD relationnelles
  - Data Warehouses: relationnel, cubes multi-dimensionnels
  - Données de transactions
  - BD orientées objets, spatiales, multimédia, textuelles
  - Données temporelles et séries temporelles
  - Données du Web
- Mais le plus souvent, pré-traitées et intégrées dans une table unique sur laquelle la recherche d'un modèle est réalisée

16

## Fonctionnalités du Data Mining

- Objectif : Décrire ou Prédire
- Caractérisation–Discrimination
  - Requêtes SQL
  - Requêtes OLAP
  - Description analytique
  - Mesures statistiques

17

## Fonctionnalités du Data Mining

- Analyse d'association (corrélation et causalité) :
  - Découvrir des règles d'association de la forme  $X \rightarrow Y$  où X et Y sont des conjonctions de termes attributs-valeurs ou prédicats
  - Mesures de *support* et *confiance* indiquent la portée et la précision de la règle
- Exemples
  - $Achat=pain$  et  $Achat=café \rightarrow Achat=beurre$   
(*support* = 5%, *confiance* = 70%)
  - $Age>20$  et  $Age<29$  et  $Revenu>1000 \rightarrow Achète\_PC="oui"$   
(*support* = 2%, *confiance* = 60%)

18

## **Fonctionnalités du Data Mining**

- Clustering
  - Trouver des groupes ou classes d'objets tels que la similarité intra-classe est élevée et la similarité inter-classes est faible
  - Pas de variable identifiant la classe
  - Classification (apprentissage) non-supervisé : classes inconnues à l'avance
- Exemples
  - Segmentation des clients
  - Cluster d'étoiles caractérisées par leur luminosité et leur température

19

## **Fonctionnalités du Data Mining**

- Classement et Prédiction (classification supervisé)
  - Apprendre un modèle qui associe un objet à une classe prédéfinie
  - Apprendre une fonction permettant de prédire la valeur d'une variable numérique
- Exemples
  - Classer des patients selon leur risque de développer une maladie en fonction de leurs symptômes
  - Évaluer le risque d'un incident de remboursement en fonction des caractéristiques des demandeurs de crédit
- Forme du résultat
  - Arbres de décision, règles de classification, réseaux de neurones, classifieurs Bayésiens, graphes d'induction, etc.

20

## **Fonctionnalités du Data Mining**

- Analyse de déviations
  - Déviation : objet qui n'est pas conforme au comportement général
  - Donnée bruitée ou exception : information utile dans le cas de détection de fraudes ou d'évènements rares
- Recherche de corrélations
  - Analyse de régression
  - Recherche de motifs séquentiels
  - Analyse statistique

21

## **Préparation des données**

- Données réelles imparfaites/endommagées
  - Incomplètes
  - Bruitées
  - Incohérentes
- Nécessité de préparer les données
  - Nettoyage
  - Intégration et transformation
  - Réduction
  - Discrétisation

22

## **Types de données**

- Numériques linéaires
  - Ex : poids, taille, longueur, vitesse, etc.
- Binaires : une valeur parmi deux possibles
  - 0 : la variable est absente, 1 : la variable est présente
- Nominales : valeur prise dans une liste finie
  - Ex : couleur peut être « vert, bleu, rouge, jaune, noir »
- Ordinales : l'ordre des valeurs est plus important
  - Ex : résultat d'un concours
- Par ratios : variables numériques sur une échelle exponentielle/logarithmique
  - Valeurs non-linéaires

23

## **Préparation des données**

- Nettoyage
  - Compléter les valeurs manquantes, lisser les données bruitées, supprimer les déviations et corriger les incohérences
- Intégration
  - Intégrer des sources de données multiples
- Transformation
  - Normaliser (ex. pour le calcul de distances)

24

## Préparation des données

- Réduction
  - Réduire le volume des données (agréger, supprimer une dimension, etc.)
- Discrétisation
  - Pour les attributs numériques, permet de réduire le volume

25

## Valeurs manquantes

- Les valeurs manquantes peuvent être codées par diverses valeurs :
  - <vide>, "0", ".", "999", "NA", " ", "?", ...
  - Il est nécessaire d'uniformiser le code
- Valeurs manquantes interdites (selon algorithme)
  - Ignorer le tuple (objet/ligne)
  - Compléter la valeur à la main
  - Utiliser une constante globale
  - Utiliser la valeur moyenne
  - Utiliser la valeur moyenne pour les exemples d'une même classe
  - Utiliser la valeur la plus probable

26

## Données bruitées

- On peut
  - Trier et partitionner (discrétiser)
  - Classifier (exceptions)
  - Appliquer un modèle de prédiction (ex : une fonction de régression)

27

## Partitionnement et Lissage

- Les valeurs triées sont réparties en largeur (distance)
  - La suite triée est partitionnée en N intervalles de même amplitude
  - Amplitude de chaque intervalle  $W = (\max - \min) / N$
  - Solution la plus simple, mais les exceptions peuvent dominer
- Les valeurs triées sont réparties en profondeur (fréquence)
  - La suite triée est partitionnée en N intervalles contenant le même nombre de valeurs

28

## Partitionnement : Exemple

- Données triées :  
4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition en profondeur :
  - Part 1 : 4, 8, 9, 15
  - Part 2 : 21, 21, 24, 25
  - Part 3 : 26, 28, 29, 34

29

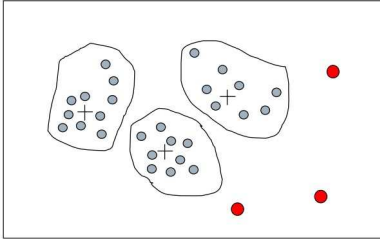
## Lissage : Exemple

- Lissage par les moyennes : chaque valeur de la partition est remplacée par la moyenne
  - Part 1 : 9, 9, 9, 9
  - Part 2 : 23, 23, 23, 23
  - Part 3 : 29, 29, 29, 29
- Lissage par les extrêmes : chaque valeur de la partition est remplacée par la valeur extrême la plus proche
  - Part 1 : 4, 4, 4, 15
  - Part 2 : 21, 21, 25, 25
  - Part 3 : 26, 26, 26, 34

30

## Classifier

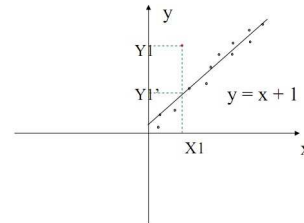
- Les valeurs similaires sont organisées en classes
- Les valeurs hors classes sont considérées comme des déviations



31

## Régression

- Les données sont lissées de manière à approcher une fonction
  - Régression linéaire
  - Régression linéaire multiple



32

## Intégration

- Combinaison de données issues de différentes sources
- Intégration de schémas
  - Identifier les entités similaires
  - Ex : id, number, matricule
- Détection et résolution de conflits
  - Résoudre les problèmes d'attributs symbolisant les mêmes entités avec des représentations différentes, des unités différentes, etc.
  - Ex : age et date de naissances

33

## Intégration : Redondances

- Données redondantes
  - Détection de données redondantes par analyse de corrélation
  - Par exemple, redondance entre attributs
  - Ex : PrixHT, PrixTTC
- Mesure la corrélation entre les attributs A et B
$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B}$$
  - $r_{A,B} > 0$  implique que A et B sont positivement corrélés
  - $r_{A,B} < 0$  implique que A et B sont négativement corrélés

34

## Transformation

- Les transformations appliquées
  - Le lissage qui supprime les données bruitées
  - L'agrégation qui calcule des sommes, des moyennes, etc.
  - La généralisation qui remonte dans une hiérarchie de concepts (taxonomie, ontologie)
  - La normalisation qui ramène les valeurs dans un intervalle donné
  - La construction d'attributs

35

## Réduction

- Permet d'obtenir une représentation réduite d'ensembles volumineux de données
- Stratégies appliquées
  - Agrégation
  - Réduction de dimensions
  - Compression
  - Discretisation

36

## Réduction de dimensions

- Suppression d'attributs : la présence d'attributs non pertinents détériore les performances des algorithmes
  - Par exemple, les algorithmes d'induction d'arbres
- Pour assurer de bonnes performances aux algorithmes d'extraction
  - Supprimer les données non pertinentes
    - Ex : identifiants des lignes, attributs à valeur unique
  - Supprimer les données redondantes
    - Données déductibles d'autres données
    - Ex : age et année de naissance

37

## Discrétisation

- Permet de réduire le nombre de valeurs d'un attribut continu en divisant le domaine de valeurs en intervalles
- Utile pour la classification, l'extraction d'associations et les arbres de décision (algorithmes qui manipulent des variables catégorielles)
- Des techniques de discrétisation peuvent être appliquées récursivement pour fournir un partitionnement hiérarchique de l'attribut

38

## Exemple

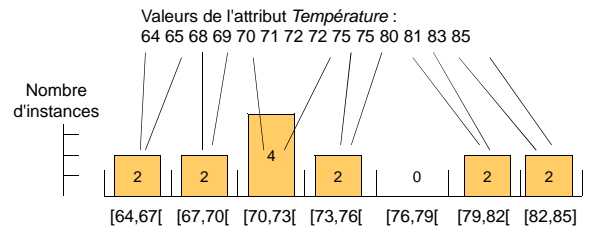
- Attribut de classe Play
  - Yes : jeu possible
  - No : jeu impossible
- Autres attributs
  - Décritent les conditions climatiques
  - Outlook : couverture
  - Temperature : degrés K.
  - Humidity : taux d'humidité
  - Windy : présence de vent

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

39

## Discrétisation en largeur

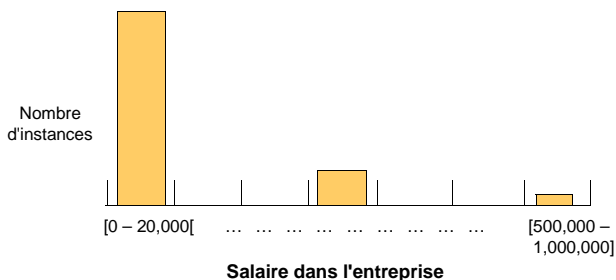
- Discrétisation par intervalles égaux des valeurs



40

## Discrétisation en largeur

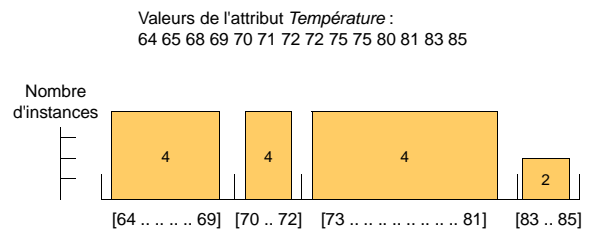
- Peut entraîner des inégalités importantes dans les effectifs



41

## Discrétisation en profondeur

- Discrétisation par effectifs égaux



- Tailles égales excepté pour le dernier intervalle

42

## Discrétisation : autres méthodes

- Présence de seuils significatifs
  - Ex : Age > 18 ans
- Discrétisation supervisée
  - Prend en compte la classification
 

temperature	64	65	68	69	70	71	72	73	74	81	83	85
class	yes	no	yes	yes	yes	no	no	yes	no	yes	yes	no
  - Utilise l'entropie pour mesurer l'information et obtenir un critère de « pureté »
 

temperature	64	65	68	69	70	71	72	73	74	81	83	85
class	yes	no	yes	yes	yes	no	no	yes	no	yes	yes	no
			no		no		yes	yes				

    - Comptage des occurrences de yes et no pour class
    - Les intervalles maximisent les co-occurrences des valeurs

43

## Exemple (2)

- Jeu de données « discrétisé »
- Valeurs catégorielles (modales) uniquement
- Simplifie l'interprétation du résultat
- Améliore les performances (temps de calcul)

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

44

## Bibliographie

- Sholom M. Weiss and Nitin Indurkha, *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann
- Michael Berry & Gordon Linoff, *Data mining: Techniques appliquées au marketing, à la vente et aux services clients*, InterEditions
- Ian Witten and Eibe Frank, *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufman

45

## Bibliographie

- Michael Berry & Gordon Linoff, *Mastering Data Mining*, John Wiley & Sons
- Jiawei Han, Micheline Kamber, *Data Mining : Concepts and Techniques*, Morgan Kaufmann
- David Hand, Heikki Mannila, Padhraic Smith, *Principles of Data Mining*, MIT Press

46

## Sites Internet

- LA référence : KDNuggets
  - <http://www.kdnuggets.com/>
- TheData Mine
  - <http://www.the-data-mine.com/>
- Conférences – Journaux
  - ACM SIGKDD – Knowledge Discovery and Data Mining
  - ACM SpecialInterestGroup <http://www.acm.org/sigkdd/>
  - DMKD <http://www.kluweronline.com/issn/1384-5810>

47

## Sites Internet

- Weka
  - <http://www.cs.waikato.ac.nz/~ml/>
- SPSS (SPSS Clementine)
  - <http://www.spss.com/SPSSBI/Clementine/>
- IBM (Intelligent Data Miner)
  - <http://www.ibm.com/Stories/1997/04/data1.html>
- SAS (Enterprise Miner)
  - <http://www.sas.com/>

48

# Sondage

